DOCUMENT RESUME

ED 296 590 FL 017 472

AUTHOR Stansfield, Charles W.; Ross, Jacqueline
TITLE A Long-Term Research Agenda for the Test of Written English.
INSTITUTION Center for Applied Linguistics, Washington, D.C.
PUB DATE [88]
NOTE 58p.; Developed through a contract with the TOEFL Research Committee.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Construct Validity; *English (Second Language); Essays; *Language Tests; *Research Needs; Scoring; Standardized Tests; Test Construction; Test Items; *Test Reliability; *Test Validity; *Writing Evaluation; Written Language
IDENTIFIERS *Test of Written English

ABSTRACT

An overview of the research needed on the new Test of Written English (TWE), a section of the Test of English as a Foreign Language (TOEFL), looks at research needs in the areas of test validity, test reliability, topic development, and equating. Suggested topics for study include: the uniqueness of the construct measured by the test, in comparison with other TOEFL scores; the comparability of scores obtained on different topics and topic types; ways in which to equate TWE scores; wording of essay test prompts; and the empirical comparability of the skills tapped by the test's two topic types. The last is considered the most important. It is noted that several aspects of the test's quality have already been well enough established that further research is not crucial. These include general validity and reliability, general construct validity, and interrater reliability. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED296590

A LONG-TERM RESEARCH AGENDA FOR THE TEST OF WRITTEN ENGLISH

Charles W. Stansfield

Center for Applied Linguistics

and

Jacqueline Ross

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

gR Tucker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL017472

INTRODUCTION

The Test of Written English (TWE) became operational in July 1986.[1] During its first year, more than 100,000 examinees took the test at three administrations; four administrations are scheduled in the 1988-89 test year. Interest in the test among university admissions officers has been strong and continues to grow. Groups other than admissions officers are interested in it as well. These include teachers of English as a second language who may be using TWE scores to place students within the writing instruction programs of English language institutes, specialists in second language testing, and specialists in composition.

The TOEFL Policy Council, the TOEFL Research Committee, the TOEFL Committee of Examiners, the TWE Core Readers Group, and TOEFL program staff agree that while the studies describing the rationale and development of the TWE (Bridgeman & Carlson, 1983; Carlson, Bridgeman, Camp & Waanders, 1985) provide ample evidence that the TWE is a valid and reliable essay test, additional research on it is desireable.[2] However, the nature of this prospective research is so diverse and the number of possible issues to be examined so large--and at times conceptually problematic--that a systematic overview is needed. For that reason, in April 1987 the TOEFL Research Committee commissioned a

project designed to develop a long-range research agenda for the TWE. The objective of this project was a paper that would outline basic issues in the determination of the validity and reliability of the test. The outline could subsequently be used to motivate and guide researchers in the conceptualization and design of TWE research projects. The research agenda could also be used by TOEFL program staff to determine the budget allocation necessary to continue the research program each year. This paper is the principal product of the project commissioned by the TOEFL Research Committee.

METHODOLOGY

The long-range research agenda presented here was developed by the individuals who served as the first and second directors of the TWE program, with input from numerous other individuals and groups. The first step in the project was to review the existing literature on the TWE, most of which is found in the reports by Bridgeman and Carlson (1983) and Carlson et al. (1985). While the external studies and comments published to date regarding the TWE are less useful in drawing conclusions, we made an effort to read them all and to cite them wherever appropriate in this paper. We also had access to ETS internal documents relevant to the TWE that were written prior to and immediately following the beginning of TWE program operations.

These included drafts of the TWE statistical analysis reports for the first three administrations, internal memoranda dealing with possible TWE equating procedures and other matters, and minutes of meetings of the core readers. In addition, we conducted a brief review of the modest body of current research literature on essay tests.

This paper describes many issues that merit research and outlines a number of studies on which work could begin soon. The issues, ideas, problems, and designs are explained in some detail in order to make the discussion accessible to the many groups that are interested in the TWE research program. Thus, in addition to providing ideas for researchers, the paper attempts to provide an overview of the research program to other potential readers, including test developers, members of the TOEFL Policy Council and its committees, the TWE Core Readers Group, TWE raters, and other interested individuals. The ideas contained herein should not be considered the only ones worthy of investigation; a long-term research agenda must remain open to new ideas and flexible enough to accommodate new concerns of the program or the field.

DESCRIPTION OF THE TWE PROGRAM

The TWE, which is administered with the TOEFL, is a direct

measure of writing ability designed to complement the indirect assessment of writing skills provided through Section 2 of the TOEFL. While Section 2, Structure and Written Expression, uses a multiple-choice format to test knowledge about written language, the TWE requires examinees to produce an organized sample of academic writing, similar to that demanded of students in many colleges and universities.

TWE examinees are given 30 minutes to write a single essay on a designated topic. Two topic types are used as writing tasks. One type, chart/graph, asks the examinee to describe and interpret a chart or a graph; the other topic type, compare/contrast, requires the examinee to compare and contrast two opposing points of view and defend a position in favor of one of them. The examinee has no prior notice about which type of topic will appear at a given TWE administration. Regardless of which task is presented, examinees are expected to address all parts of the writing question, to compose clearly in standard written English, to organize their ideas, and to support their ideas with examples or evidence. Examinees are permitted to make notes and to organize their essays in the work space provided on the Essay Page before beginning to write.

Examinee responses are scored at a centralized reading in the San Francisco Bay area within two weeks of the test date. They are scored holistically on a six-point criterion-referenced

scale by trained raters, most of whom are teachers of English composition or English as a second language. Each paper is read by two raters; if the scores differ by more than one point, a third reading is required. Off-topic responses are not rated on the scale; instead, the word OFF is printed on the examinee's score report. Examinees who fail to respond are given a score of 1, which is the lowest point on the TWE scale, along with an indication that they failed to respond. The TWE score does not contribute to the TOEFL total score, but is listed separately on the TOEFL score report for the four TOEFL administrations that include the TWE (Stansfield & Webster, 1986).

THE RESEARCH AGENDA

This section of the paper is organized into four general areas: validity, reliability, topic development, and equating. While topic development might be appropriately placed within the framework of validity, test researchers currently do very little research on item format or item writing concerns. Topic development is presented here as a separate area because of this, and because the variables of concern are not ones about which psychometricians are ordinarily informed. Quite possibly, the initiative for such research will have to come from test

developers. Equating might appropriately be placed within the framework of reliability, since it involves an adjustment for a lack of comparability in raw scores. However, because equating has become a field unto itself within psychometric science, it is considered separately here. The discussion of reliability includes studies of the essay reading procedures and the characteristics of the readers themselves. Perhaps these matters should also be considered separate areas of research. However, we have included them under reliability, because readers traditionally have been the source of data for studies of essay test reliability.

VALIDITY

In large-scale testing programs, validity is normally the principal source of concern. More than being a characteristic of the test itself, the concept refers to the inferences made about a test score, i.e., the degree to which it is useful as a measure of a particular trait for a particular purpose and for a particular examinee. Traditionally, validity has been broken down into three types: construct validity, criterion-related validity, and content validity. Criterion-related validity is often subdivided into concurrent validity and predictive validity. More recently, all forms of test validity have been viewed as aspects of construct validity, because they all provide

evidence with which to judge the utility of the test as a measure of the trait being assessed (Messick, 1987). While we agree with this view, we will use the traditional nomenclature as a convenient means of subdividing and organizing into groups the totality of validity concerns. It is important to remember, however, that all validity studies relate to construct validity.

Construct validity

The traditional question in construct validity is: "Is there evidence that the test is measuring a particular psychological trait?" As has been noted by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985), the trait should be presented as part of a conceptual framework. The framework specifies the meaning of the construct, distinguishes it from other constructs in the framework, and indicates how measures of the construct should relate to other variables. Within the realm of language proficiency, this means that each score associated with the TOEFL should measure a different construct. While constructs within a framework can be interrelated, there must be evidence that they are not identical, and the more distinct the construct, the greater the support for reporting a separate score for it.

Evidence for the construct validity of the TWE was presented

in the study by Carlson et al. (1985). This study found that the TWE topic types (compare/contrast and chart graph) show similar patterns of correlation with TOEFL section and total scores. It also found that different compare/contrast topics do not correlate more highly with each other than they correlate with chart/graph topics. The findings of this study were taken into account by TOEFL program staff when the decision was made to begin using both topic types alternately on an experimental basis. Despite the positive nature of these experimental findings, a number of other studies utilizing operational program data are warranted.

The relationship between TWE and TOEFL Section 2 scores should continue to be examined in order to gain a better understanding of what each measure is testing. Carlson et al. (1985) found that the correlation between the two is about .70, with different topics showing minor variations in the magnitude of the correlation. The factor analysis of TWE and TOEFL scores they carried out showed that each loaded on separate factors. They suggested that the differences could be attributed to method of testing. However, as they noted, one might also conclude that the former probably taps a production factor while the latter probably taps a recognition factor. This suggests that the two measures should continue to be reported separately, since there is enough unique variance associated with each to assume that each represents a separate construct. However, this assumption

should be examined further. If it were established that the two were measures of the same construct, it might be possible to use Section 2 scores to statistically equate TWE topics, to combine both sections of the test, or to eliminate one of the measures entirely. The practical payoff from such a finding could be enormous. However, even if research showed that the two subtests measured the same construct, one would have to consider the political ramifications of the latter two actions. Since a major reason for developing the TWE was the perception among specialists in the ESL field that a direct measure of writing was needed, it might not be wise to eliminate the TWE or even to combine it with Section 2.

A fundamental construct validity concern that is also related to reliability is the comparability of the two topic types currently used on the test. Carlson et al. (1985) found that the compare/contrast topic type and the chart/graph topic type produced similar means and standard deviations in their sample of TOEFL examinees and that the examinees were ranked similarly on both topic types. This suggests that the <u>type of topic</u> an examinee is given does not make a difference in the score (although, due to less than perfect reliability, there is a topic effect in that examinees perform differently on different topics). The two topic types also correlated similarly with TOEFL section and total scores, which suggests that the two tap the same construct. Although similar means, standard deviations,

and correlations with other measures would normally be considered strong evidence that both topic types tap the same construct, there are several reasons why the matter deserves further examination.

First, since the findings of Carlson et al. (1985) were obtained using an experimental version of the TWE, and the wording and design of the typical chart/graph topic has evolved slightly as TWE staff and consultants have gained more experience, this change in topic format could be affecting examinee performance. There seems to be some concern that this might be the case. At recent meetings, the TWE core readers have expressed the opinion that the chart/graph topics tap different writing skills than the compare/contrast topics. If the opinion is valid, the two topic types would not be interchangeable. In a review of the studies that led to the TWE, Greenberg (1986) expressed doubts that the two topic types measure identical constructs, and Roy (1987) in an informal survey has noted that at least one researcher she spoke with also expressed doubts that the two are comparable. Thus, the comparability of the two topic types should be given further study through appropriate statistical analyses of examinee performance. One approach would be to replicate the Carlson et al. study using operational topics. However, this approach may be sensitive to the unique characteristics of the topics selected to represent each topic type. Another, perhaps more powerful, approach would be to group

statistics for all topics of each type and examine them for differences in mean, variance, and correlation with TOEFL scores. After many topics of each type have been administered, it may be possible to reach some generalizable conclusions about construct validity across topic types. If further research indicates that the two topic types are testing the same construct, one might also conclude that a principal difference between them is in the cognitive and linguistic skills that readers perceive each to measure. Evidence of this difference in reader perception is found in Carlson et al. (1985).

Because at present the TWE addresses only two kinds of academic writing tasks, and many other kinds of tasks are used in academic settings, it would be desireable to expand the number of topic types available for a given TWE administration. Therefore, research should be carried out on the comparability of other topic types to the two presently included. A cause/effect topic type or an argument-to-a-designated-audience topic type should be investigated since these require cognitive and organizational skills that teachers of rhetoric find attractive. The sequential/chronological description that Bridgeman and Carlson (1983) found acceptable to faculty in all disciplines should also be studied.

The construct validity of the TWE can only be assessed directly through statistical analyses of the psychometric characteristics of the topic types employed on different forms of the test. Yet some writing specialists we have contacted believe that linguistic analyses of examinee writing also have a bearing on the issue of whether the two topic types assess the same construct. In this sense, a number of studies should be conducted that examine and describe different aspects of examinees' writing. However, while linguistic analyses would be useful to score users, in that they would describe the written English produced by examinees at each score level and on each topic type, they would not directly address the question of the

comparability of the constructs being measured. The dilemma can best be described by referring to another second language skill, such as speaking. One might assume that engaging in everyday conversation or dialogue with professional associates calls for fundamentally different skills than lecturing in one's profession. This is, of course, a testable hypothesis, from both an empirical and a qualitative standpoint. However, if an empirical analysis of performance ratings on each task failed to show any significant variation between the two sets of ratings, one could claim that both tasks were tapping the same construct, even if qualitative differences were found in the language used on each task. Thus, we see that two tests that appear to measure different constructs may actually be measuring the same construct. Again, the results of research should be useful.

Criterion-related validity

Criterion-related validity refers to the degree to which test scores are related to scores on other measures. These measures may be more established tests of the same construct, or they may be indices of the actual outcome variables (knowledge, skills, and abilities) the test purports to measure. The measures with which such a test is compared are called criterion variables. When scores on the criterion variables and the test being validated are obtained at about the same time, they are evidence of concurrent validity. When scores on the criterion variables are obtained substantially later, they are evidence of predictive validity. Traditionally, predictive validity has been considered relevant to selection tests. However, since language proficiency is a construct that is open to intervention through instruction or other experiences, studies of concurrent validity are probably more relevant to TOEFL program tests such as the TWE since they are less likely to be contaminated by the effects of such intervention. In general, care must be exercised in the selection of a criterion variable. Otherwise, one may reach misleading or even false conclusions regarding the validity of a test. Often, it is necessary to establish the reliability and validity of a criterion variable through research before using it in a study of criterion-related validity. While this is most often the case with criteria that are surrogate measures (e.g.,

faculty perceptions of student competence), research may be necessary to justify the use of other types of criterion variables as well.

Concurrent validity with direct measures

Criterion-related validity could be assessed by administering the TWE to a group of foreign students enrolled in regular university classes and examining the relationship between their TWE scores and scores on a portfolio of different classroom assignments. Such assignments might include essays or compositions written under timed conditions in an English class, compositions written as overnight assignments, term papers in content courses, lab reports, book reports, summaries of articles, and so forth. One could examine the relationship with grades on each type of assignment, and the relationship with faculty ratings of the adequacy of the student's English writing skills for handling such assignments. Faculty ratings, while not without shortcomings as a criterion variable, might produce higher correlations with TWE scores than would grades, since grades would be influenced by confounding variables such as the examinees background in the field of study. Still, both types of criterion-related validity should be examined.

Concurrent validity with surrogate measures

One would also want to know the relationship between TWE scores and university post-admissions essay placement tests. These would serve as variables for establishing concurrent validity with other, perhaps more established, measures. In justifying such a study, one might assume that such measures might be closely grounded in institutional writing programs and therefore be quite valid. If the TWE were correlated with such measures, the correlation coefficient would demonstrate the degree to which both measures tap the same construct. In certain cases, institutions may want to compare the criterion-related validity of their own post-admissions assessment and the TWE with performance on actual academic writing assignments. Such studies would typically come to the attention of the TOEFL Program Office as an external request for research support. It would be desirable to provide assistance to such studies, because the outcome would also allow ETS to make a comparative evaluation of the criterion-related validity of the TWE.

Predictive validity

Predictive validity of the TWE could be assessed by collecting and analyzing faculty ratings of TWE examinees after they had enrolled in regular university settings. In this case, the TWE scores would typically have been obtained about one year

prior to the study. In such a study, it would be useful to exclude from the sample those students who had received additional writing instruction (in an English language institute or a regular university composition course) after taking the TWE, since this instruction might also affect their performance on writing assignments and, subsequently, faculty perceptions of their competency.

Test bias

There are a number of popular and theoretical issues related to test bias that represent threats to the validity of the TWE. Test bias issues should be investigated as concern for them arises among the public or within the research community. While freedom from test bias is not a type of validity, the absence of an irrelevant bias is surely part of validity. Test bias is discussed here within the framework of validity, because the concern has implications for perceptions of validity and for user interpretations of the validity of scores for different types of examinees.

One relevant current concern among second language writing specialists pertains to assumptions about what is often referred to as contrastive rhetoric. According to Kaplan (1972), different cultural groups have distinctly different ways of

organizing their ideas in written discourse. Citing examples of Western, Oriental, and Arabic rhetorical patterns, he claims that these patterns persist when non-Western people write in English. If this is the case, it would be useful to know if such patterns can be identified on TWE essays and whether they have any effect on TWE scores. Although some people would argue that such a rhetorical pattern effect on TWE scores would represent an undesirable cultural bias in the test, others would argue that such scores would represent a valid assessment of the examinee's ability to communicate in written English. Regardless of which position is correct, given public concerns about bias and the credibility the hypothesis enjoys within segments of the ESL profession, it would be useful to determine if such patterns appear systematically among the responses to TWE writing tasks, and whether such patterns seem to play a role in the TWE score. The TOEFL Research Committee has funded a proposal by Freedle (1985, September) designed to address the issue.

It might be desirable to conduct other studies that focus on the test bias issue. Consideration should be given to the effect of examinee background variables on test performance. In particular, given traditional public concerns about the validity of standardized tests, one might want to examine the effects of sex, race, cultural background, native language, native country, major, and other variables on performance. Such studies might be able to make use of the Mantel-Haenszel index (Mantel & Haenszel,

1959), which matches two groups of examinees (e.g., male versus female, Oriental versus others, Spanish speaking versus others) on an ability, such as English language proficiency, and then produces an index of the degree of discrepancy in the dependent variable, which is the score on another measure. In this case, the dependent variable of interest would be the TWE score. Theoretically, if no biases exist, no systematic differences in the TWE scores of the two groups should be found.

One obstacle to an adequate design in such studies is determining what will be the valid matching variable. In theory, this variable should be another measure of the same construct. Thus, it is not entirely appropriate to match on overall TOEFL score. However, since the TWE is a one-topic test, no other direct writing score is available for matching. One solution might be to match groups on the Section 2 score (Structure and Written Expression). However, as noted earlier, this section may represent a different construct (knowledge of conventions of writing as opposed to ability to write). Despite the problems associated with attaining a fully adequate design to address concerns about bias, the matter of bias merits attention. While one might not be able to draw firm conclusions from the outcome of preliminary studies with less than satisfactory designs, such studies might provide informal evidence that a more tightly controlled study of topic bias is warranted. For instance, if, after matching on Section 2 score, it was noted that a group

consistently performed lower on a particular topic type, it might be desirable to carry out a large-scale experimental study in which examinees write on more than one topic type and then examine the differences among different native language groups.

Again, in considering the question of bias, one must remember that differences in performance do not necessarily indicate that a topic or test is flawed. Such performance differences may be valid indicators of true differences between groups in terms of their ability to deal with the task represented by the topic. To determine that a topic is inherently flawed in a way that unfairly disadvantages certain groups apart from their ability, a qualitative analysis focusing on the content of the topic may be necessary.

Content validity

Bridgeman and Carlson (1983) provide information that supports the content validity of the TWE. They found that foreign students in North America are indeed asked to compare and contrast two different points of view and take a position in favor of one on assignments given in a wide variety of fields and that, similarly, foreign students in the social sciences, natural sciences, engineering, and business are asked to describe and interpret charts and graphs. This suggests that the topic types

used on the TWE thus far are similar to the writing tasks required in an academic setting. However, Bridgeman and Carlson investigated the frequency of use and faculty perceptions regarding only ten academic writing tasks. Other writing tasks may also be used and further research should be carried out to determine what they are and how faculty perceive their validity as a measure of academic writing. Of course, the content validity of the test would be increased if it included topic types representing a larger proportion of those tasks actually used in academic settings. If faculty indicated that other writing tasks are important, those tasks should be incorporated into the TWE program. Of course research on topic comparability will have to be carried out before this can occur.

In a multiple-choice test, the test items contain all the content features that produce statistical differentiation--in particular, item difficulty and item discrimination power. In an essay test, such characteristics do not seem as relevant. Indeed, in the writing of the topic, an effort is made to construct a prompt that will not be difficult, and will be accessible to the widest possible range of students. Discrimination power is relevant on an essay test, for without it there would be no differentiation among examinees. However, differentiation among examinees is provided by the essay readers, who must identify different levels of writing ability based on the examinees' essays. Thus, we see that both the examinees'

responses and the way the raters react to those responses constitute much of the data of concern to psychometricians and other professionals in an essay test program.

Similar fundamental differences exist when judging the content validity of an essay test. While the content validity of an essay test may be judged by evaluating the similarity between essay prompts and real-life writing tasks, it can also be judged by examining a number of essays to see if they appear to represent the kind of performance skills examinees will need subsequently. While content-related, face validity judgments can be made by "eyeballing" the essays; more systematic study involving linguistic analyses of the essays is required in order to provide information that can be used to judge content validity. This means that a second type of content validity study will involve analyzing and describing the writing elicited by topics at each score level in order to provide the score user with an objective description of what the test measures. Such analyses and descriptions should be carried out by score level, by topic, and by topic type.

Such studies would also relate to issues other than content validity by providing a better understanding of what TWE scores mean. Score interpretation would be enhanced by linguistic analyses that contributed to a more complete understanding of TWE topic types in terms of the comparability of the linguistic

skills required to respond to them. Linguistic analyses would provide information that could be used to identify the types of language associated with specific topics and the differences in language used in response to different topics. Further statistical analyses could be carried out on the results to determine their relationship to examinee background variables, such as native language, country of origin, sex, and degree sought. Finally, such linguistic data can be combined with TWE scores and analyzed statistically to determine the linguistic and rhetorical correlates of specific TWE scores.

Some preliminary studies of this nature have already been carried out. As part of the study by Carlson et al. (1985), Joy Reid of Colorado State University analyzed about 80 essays written on each of four topics--two compare/contrast topics and two chart/graph. The analysis involved using the text analysis criteria of the Writer's Workbench developed by Bell Laboratories. Reid found that the criteria of number of words, number of content words, number of short sentences, and number of "to be" verbs showed the most significant and consistent relationship with essay ratings. Other variables, such as number of long sentences, percentage of passives, percentage of prepositions, and percentage of conjunctions also were related, although the relationship varied across topics and topic types, suggesting that different topics and types require different linguistic features to communicate in writing. As noted in the

Carlson et al. study, the fact that different linguistic features are associated with different topics suggests that the readers are able to adjust their standards to different topics and modes of discourse.

Conner (1987) analyzed a sample of 22 essays written by her students at Indiana University in response to a disclosed compare/contrast topic from the TWE operational program. Using three rhetorical features of persuasive writing (claim, data, and warrant) developed by Toulmin (1958), she found fairly good correlations (.68 - .72) between TWE score and the number of such features in each essay. These correlations are higher than those obtained by Reid (see above) using linguistic variables, suggesting that the examinee's ability to utilize certain rhetorical features plays a role in TWE scores. Studies using larger samples and actual TWE examinees and raters would be more informative. Other rhetorical features may also play a role in the ratings, and the presence and role of these features may vary according to topic type. For instance, as noted by Conner, one might examine whether other rhetorical modes (chart/graph, etc.) elicit the same features of persuasion as a compare, contrast, and take a position topic.

Another desirable focus for linguistic and rhetorical analyses is discrepant essay papers. A paper that receives ratings that differ by two or more points may be the victim of

random error in one or both ratings. On the other hand, such papers may exhibit special combinations of linguistic and rhetorical characteristics that are difficult to evaluate consistently. Studies of such papers should be carried out to give us a better understanding of this anomaly.

Another useful linguistic/rhetorical analysis would be a comprehensive study of writing at TWE level 4. As described on the scoring guide, 4 seems to be the point that reflects basic writing competence while scores lower than 4 demonstrate a lack of it. What can an examinee with a score of 4 do with the written language? What problems does he or she still exhibit? What problems does he or she exhibit on concurrent classroom writing tasks? What remediation is suggested by these problems? What are the curricular implications of these problems? A study addressing such questions would be very useful for English departments and English language institutes, and it would address the goal of the Research Committee and the Committee of Examiners to provide more diagnostic information in vital skill areas to examinees and score users. While analyses of language skills at each level would be useful, the first analysis should focus on level 4.

Survey of examinees

A formal survey of TWE examinees might provide useful information concerning various aspects of the test's validity and other matters. Stansfield (1986b) identified some examinee perceptions in a discussion with students at the University of the District of Columbia test center following the first TWE. He reported that about 80% said they had adequate time to answer the question, most were positive about the inclusion of a writing sample, and all preferred to complete the TWE before the regular TOEFL. For a more representative sample of examinees, one would want to know their perceptions of the validity of the essay prompt and the conditions under which it was administered (time allotment, administration before versus after TOEFL), their prior expectations, preparation behaviors, reason for taking the test, and so forth. The survey could also inquire into examinees' cognitive strategies for responding to the TWE prompt by asking questions such as "Did you make notes before beginning to write?, How long did you spend making notes before beginning to write?, Did you edit/check your response after writing?," etc. Such a survey questionnaire could be completed either immediately after taking the TWE, or it could be given to examinees to complete at home and return by mail. Of course, the former approach would be more satisfactory, since it would ensure the participation of a random sample of examinees in the study.

RELIABILITY

As indicated in the introductory discussion of validity, in most testing programs validity is the major research concern. However, in essay testing, reliability is of greater than normal importance. It may even be a more pervasive concern to the test publisher than validity. This is because essay tests exhibit a good deal of face validity since they require the examinee to perform instead of demonstrating knowledge about how to perform. Essay tests pose special reliability problems because they are open to sources of error that are not present in multiple-choice tests. Since the credibility that accrues to direct measures is useless without reliability, reliability becomes a pervasive concern.

Traditionally, reliability has been divided into several categories, each of which relates to the source of the error of measurement. One kind, test-retest reliability, refers to the consistency of the examinee's performance at different administrations. The source of measurement error here is the examinee or, more specifically, those factors that may cause his or her responses to vary from day to day. A very similar type of reliability, parallel-form reliability, refers to consistency of performance on different versions of the test. Here the measurement error emanates from the test itself, since the items on any form of the test are an incomplete sample of the universe of items that could be included. Essay tests may exhibit both types of error, but a third source of error, the rater, is

usually the source of greatest concern. The consistency of performance by a single rater is referred to as intra-rater reliability. Consistency of performance across raters is referred to as inter-rater reliability.

Inter-rater reliability

The results obtained through statistical analysis of the TWE forms administered thus far indicate that the inter-rater reliability of the TWE is outstanding, in comparison that found in other essay testing programs for which statistics are available. A lower-bound estimate of the reliability of a single TWE rating is usually about .75, and the reliability of the score based on two ratings is usually about .86. While inter-rater reliability is normally reported as a measure of the reliability of an essay test, it is important to remember that such an index is only based on one of three sources of essay test error. While it is possible to calculate the inter-rater reliability based on data obtained at each reading, it is necessary to implement an experiment to determine other types of reliability. If the TOEFL program is to have a comprehensive picture of the reliability of the TWE, such experiments will have to be carried out.

Test-retest reliability

This type of reliability is determined by giving the same test to the same people at different times and then correlating the two sets of scores. However, for several reasons, it is often impractical to determine test-retest reliability due to the

effects of test "reactivity." Reactivity is a phenomenon in which the first administration sensitizes the examinee to the task being measured, with the result that the performance on the second administration will be different. In the case of the TWE, the first sitting could serve to instruct the examinee in the problem presented. It is likely that, having thought out the problem beforehand, the examinee could improve on the presentation on the second sitting, just as one improves on sequential drafts of a report. Another reactivity problem in determining test-retest reliability is memory. Instead of approaching the task from a fresh perspective, the examinee may simply attempt to recall the previous response and put it in writing, and this reaction would give an inflated estimated of the true test-retest reliability.

Parallel-form reliability

A common solution to the inability to determine test-retest reliability is the reporting of parallel-form reliability. Similar to the test-retest method, it differs in that different forms of the test are used in subsequent administrations. Although this solution is motivated by practical concerns, it is important to remember that parallel-form reliability reflects error in both the examinee and the form. It is probably more useful to approach the reliability of the TWE in this way,

because in practice, the examinee would receive a different form of the test at each administration, and therefore both sources of error would contribute to variation in the score.

One way of determining parallel-form reliability is through a design whereby examinees are asked to write essays in response to different prompts on different days, while controlling for the order in which prompts are given. The interval between prompts could be between one and seven days. It might not be useful to have a longer interval because that might allow for a change in actual writing proficiency in the interim. A serious concern in such a design would be how to control for motivation to perform well on both forms of the test, since the examinee would know which form was the operational form and might be more inclined to respond carefully and thoughtfully on the operational form. One would also want to control for the order in which prompts are given.

It would also be interesting to know if there would be an appreciable effect on TWE scores if test time were lengthened. This could be determined by giving a paid sample of examinees two essays, e.g., a 30-minute essay followed by a 45-minute essay. To control for topic comparability, one could make the second essay the same essay that had been given earlier to examinees in another part of the world. Thus, one could predict the score on the second essay based on the TOEFL scores and the score on the

first TWE essay. Similarly, one could create an even stronger design (one with a concomitant replication) by using two such groups. For instance, examinees in zone B would get the topics for zone A (with additional time) and zone B (regular time), and candidates in zone C would get the topics for zone B (with additional time) and zone C (regular time).

The design could be strengthened even further by telling examinees that both essays would be counted in their score. This approach to TWE research would generally seem to be the strongest in terms of generalizability. The cost of administration would be higher since the test would last twice as long; however, the cost of scoring could be held constant by rating each of the two essays only once. Another practical implication of conducting this research within the framework of the TWE operational program would be that, for the particular administration involved, the reliability coefficient obtained would be an index of parallel-form reliability rather than the inter-rater reliability currently reported. One could obtain inter-rater reliability in addition by simply doing a second rating on a sample of essays. In studies comparing different topics and topic types, a design that involves the administration of two topics should be given serious consideration.

Another, even stronger design, would be to randomly assign testing conditions (test topic and test length) to examinees at a

particular administration. Thus, examinees at each test center would receive all topics, but these topics would be spiraled so that each examinee would only have to write on a single topic. By randomly assigning topics in this way, it would be possible to do an analysis of variance by topic. While such a design might pose concerns about topic security, in that examinees in some parts of the world might have time to learn of the topics presented in a time zone that took the test earlier on the same day, this would only be a problem if topics were spiraled regularly. If this were carried out as an unannounced experiment, there would be no reason for examinees to try to obtain the topics presented in the earlier time zone, since ordinarily the topics vary by time zone.

Currently, the TWE does not allow a choice of topic due to a concern that this could have a variety of negative effects on examinee performance. It would be helpful to know the effect of allowing a choice of topic on performance. In a study of the question involving ESL learners at an American university, Leonhardt (1985) did not find that choice of topic had a significant effect on performance. However, many essay tests allow a choice of topic, including most ESL admissions tests produced in Britian and the Hong Kong Examinations Authority school leaving examinations. The issue could be assessed by permitting all examinees at certain test centers to choose between two topics: one regular and one from a previous time

zone as in the design described above. Examinees would first be told to choose and write on the topic they felt most capable of handling. Then, after one half hour, they would be told to write on the other. Because they would not know which topic was the operational topic for their zone, motivation to perform well would be controlled. Such a design would also permit an assessment of parallel-form reliability using an operational population and operational topics. The Carlson et al. (1985) study provided a measure of parallel-form reliability with a real TOEFL population. However, the topics they used were not true operational topics, and there are minor differences between their prompts and the nature of both types of prompts that have appeared on the operational TWE so far.

Another approach to assessing the effect of choice might be to randomly assign choice and no-choice conditions and determine the correlation between the two conditions as well as the effect on the mean and variance of scores.

As indicated at the beginning of this section, essay tests involve at least three basic sources of error: the examinee, the rater, and the topic. These multiple sources of error can explicitly be taken into account by using an approach based on generalizability theory. By including the sources of error as factors in a factorial analysis of variance design, in a generalizability study one can determine the effects of each

factor and interactions of factors on the total variance. In the field of second language research, its usefulness has been discussed by Bolus, Hinofotis, and Bailey (1982), and it has been employed by Brown and Bailey (1984) to assess the components of total variance on a single-topic essay test for ESL learners. Generalizability theory should be considered a potential method of analysis for obtaining a more accurate picture of the reliability of the TWE.

Studies of rating procedures

In a subjectively scored test, an examinee's score is determined both by the written response of the examinee and the reaction of the rater. An area of investigation that is indirectly related to test reliability, and directly related to the improvement of operational procedures, is research on the training and qualifying of raters. The TWE program is one of only a few ETS programs that require new raters to qualify (demonstrate adequate reliability and speed with essays) before being invited to a reading. While this policy may be partly responsible for the high degree of inter-rater agreement achieved at TWE readings, it would be valuable to know to what degree further gains can be made. If one considers the qualifying task (uninterrupted rating of 50 essays following training) as a test, it would be possible to examine the predictive validity of the

qualification program by correlating it with each rater's reliability and speed at the first reading. Generally, one would attribute less importance to speed than to reliability. However, by indicating how long it takes readers to complete the qualifying task and then correlating it with average number of essays read in the rater's first operational reading, one could also determine the predictive validity of this aspect of the qualifying process.

Change in rater performance over time is also of interest. It is possible that raters may become more adept at reading, but it is also possible that they may become more complacent readers, thereby declining in quality with experience. Thus, it would be useful to routinely examine raters' reliability and speed statistics in order to determine typical patterns of growth or attrition. This would give program direction and essay reading staff a better understanding of the reading process.

Individual components of the reader training procedures used at the outset of a reading may also be evaluated through the use of experimental designs. By randomly assigning raters to two different reader training procedures, one could determine if either procedure results in improved reliability or a difference in scores of examinees who wrote on the same topic. This might be one way of assessing whether it is permissible to place raters (of the same prompt) in different rooms with different assistant

chief readers when a large enough room to accommodate all is not readily available.

Another type of study that could be incorporated into reader training is ethnographic, process-oriented research. Under this approach, a specialist in ethnographic research methodology could participate in the reading as an observer and make notes on significant events that occurred during the reading and their apparent effect on readers. The researcher might also survey or interview readers subsequent to the reading to follow up directly on the effect of the event. Such research might produce insights into the effect of such things as specific examples used in training, specific questions asked by readers (and the response of the chief reader), group discussions, and the table leader's behavior toward readers (praise, criticism, etc.).

Studies of raters

One could also conduct background studies on the raters, just as studies of examinee background characteristics are typically conducted in large scale, multiple-choice test programs (e.g., Wilson, 1982). One outcome might be the development of profiles of typical readers of different types. Another outcome of such studies might be the documentation of characteristics of good readers. Such studies could identify and quantify

significant relationships between rater background characteristics and performance as a reader. Multiple regression equations based on reader characteristics might make it possible to objectively screen applicants for training in order to identify those with a high probability of success. Among the obvious reader characteristics that might be included are age, profession, years of teaching experience, essay reading experience, field of teaching, foreign language background, experience living abroad, and frequency of interaction with foreigners. The outcome of such studies might even challenge traditional beliefs about the characteristics of good readers.

It might also be possible to conduct studies of readers' cognitive and personality characteristics and their relation to reading speed, reliability, and consistency over time. For instance, one might posit that a positive relationship exists between reader reliability and field independence, which is the ability to disambiguate or perceive order in a seemingly ambiguous field of information. A field-independent reader should be readily able to disambiguate (understand) an essay written by a nonnative English speaker.

TOPIC DEVELOPMENT

Research should also be carried out on topic development concerns. While ideally only the examinee's ability should dictate scores, practical experience indicates that both the reader and the prompt affect scores as well. Within this paper, we will leave the discussion of differences between topic types to the section on validity. Thus, we will focus our discussion of topic development issues on those aspects of the wording of the prompt that may affect the examinee's response. It is important to remember that the response can be affected qualitatively (in terms of its linguistic and rhetorical characteristics), psychometrically (in terms of ratings), or both. We will begin with a discussion of the compare, contrast and take a position type of topic.

The Core Readers Group and professionals within the composition field at large, have considerable experience in designing topics of the compare, contrast, and take a position type. This experience has produced a sense of the topics that engage an examinee, are accessible across a range of ability levels, and produce a variety of responses. The current members of the Core Readers Group believe that minor differences in wording may affect performance because they can affect the literal meaning of the rhetorical specifications of the task. Some of these differences are inherent to the content. For instance, if a prompt states a scenario (such as the building of a large factory in one's home town), the examinee may be asked to

compare the advantages and disadvantages of the scenario. The same prompt can be worded differently, for example, by adding that some people support this proposal while others are opposed--and the examinee can be asked to compare both points of view. Although the basic content of the scenario is the same in each prompt, differences in the focus of the prompt could potentially result in differences in the ways examinees address the writing task.

Another compare/contrast prompt may put forth two possible actions and ask examinees to compare them and indicate which they prefer. Although still a compare/contrast prompt, the writing task in this case is clearly different from that of the previously mentioned prompts. Such a prompt was used by Carlson et al. (1985) in a question pertaining to active versus passive leisure activities. One method of controlling such variations in wording would be to allow for only one type of prompt in the specifications. If this produced more consistent results it might be desirable. However, tightly controlled task specifications reduce the validity of the test, make it more coachable, and restrict the creativity of topic developers. Research about the effect of such variation in the task specifications on scores might produce more information for decision-making.

Another possible problem relates to the specificity of the

instructions contained in each prompt. Typically, examinees are told to give reasons to support their answers, and sometimes they are told to explain their choices. Such specific instructions may reduce variability in the scores, since they may help the examinee organize the answer. This added guidance might assist low- and middle-level writers who might otherwise attain lower scores. It might also contribute to more similar responses, which in turn may contribute to reader boredom, which in turn may reduce reliability. Perhaps one should leave the organization of the essay wholly to the examinee. The matter could be researched. If an effect is found, it could influence the design of prompts in the future.

While the wording of chart/graph topics is also important, this topic type itself presents more fundamental questions. A basic concern is that the charts organize the evidence for the examinee, and provide the basic vocabulary for discussing that evidence. As a result, there is a tendency for such prompts to elicit more systematically structured responses than compare/contrast prompts do. This can affect the mean and reduce the variability in scores. However, the tendency can be more marked with some chart/graph topics than with others. It may be possible through research to gain an understanding of which characteristics of a chart/graph topic may facilitate responses, thereby affecting the mean and variability and, therefore, score comparability.

Pictorial detail is also a concern when developing chart/graph topics. Too much detail can confuse an examinee who is inexperienced at reading such material. Too little detail does not provide enough information to elicit writing from which to judge the examinee's writing ability. The relevant variables involved could be identified and researched.

Wording is also of considerable importance with chart/graph topics, because it shapes the rhetorical structure of the response. At present, some of the chart/graph prompts call for a response that is basically descriptive; others call for suasion. One could investigate the effect of specifying that the response should involve persuasion in addition to description. While research should be carried out on the effects of specific types of wording, it may be that the effects are mainly on the structure of rhetoric produced by the examinee, and not on the score itself. While raters may be able to compensate for minor qualitative differences in responses when using the scale, it is important to the program that raters, who are themselves teachers of writing, be satisfied with the samples of writing elicited by TWE prompts. Research on the wording of prompts may improve satisfaction with the type of response elicited.

EQUATING

Equating requires adjusting for the difficulty of a measure based on the average performance of the same or an equivalent group on another measure. Because only one topic is given on the TWE, it is not possible to equate the essay score to performance on another topic.

On different occasions, the core readers have expressed the fear that it is easier for examinees with limited English language proficiency to perform well on chart/graph topics than on compare/contrast topics. It is also noteworthy that there is some evidence that this is the case based on data contained in the statistical analysis reports issued for TWE topics to date. However, because differences can be found in the average language proficiency (as measured by the multiple-choice TOEFL) associated with the mean TWE score on topics of the same type, additional evidence will have to be gathered before conclusions can be drawn.

One way to determine whether differences exist in the difficulty of topics and topic types is through an experiment in which at least six test topics (three of each type) would be randomly assigned within all test centers. Because the groups taking each form would be equivalent, differences in performance could be attributed to differences in the difficulty of topics

and topic types. If significant differences in the difficulty of topics were found, it might be necessary to adjust statistically for it. One way to do this would be to continue to spiral several topics within each test center at each administration. However, this would place considerable strain on the test development process, since it would require additional item writing, pretesting, and pretest reading, and it would increase the number of topics to be read at each reading. If, in the above mentioned experiment, it turned out that significant differences were found only in the difficulty of topic types, it might be desireable to take some action involving a new policy, such as informing score users about the nature of the trend and indicating the type of topic assigned on the score report. Such an experiment, which might be done as part of a construct validity study, merits fairly high priority.

When the TWE program was being designed, it was proposed to equate TWE scores by linking them with the score on TOEFL Section 2, which is believed to be an indirect measure of writing skills. However, this approach assumes that Section 2 and the TWE measure the same constructs, an assumption that is contradicted by Carlson et al. (1985). Therefore, it was decided to report the TWE score separately, without using TOEFL section scores for statistically equating different prompts, and to anchor it in a carefully derived, performance-based scale, the development of which was described by Stansfield (1986a). Many additional

procedures are used to maintain the consistency of scoring standards at TWE essay readings. These include having the chief reader, assistant chief readers, and table leaders reread the benchmarks used at the previous reading before selecting benchmarks for the current reading, carefully training readers on the benchmarks and scoring guide, having table leaders continuously review essays read by each rater to confirm the rater's adherence to the scale, and so forth. Taken together, the scoring guide and the extensive quality control procedures imposed prior to and at TWE readings constitute the TWE program's method of equating. Still, the possibility remains that these procedures do not fully ensure score comparability across TWE forms. The matter deserves continued attention.

Another possible way to equate TWE scores with the data currently available is to examine current TOEFL items in an effort to identify a subset of items that consistently correlates highly with the TWE score. It may be that a subset of items (whose content specifications can be identified) from one, two or all three sections can be used. If it were determined that these items load on the same factor as the essay, they would provide a relatively error-free link between the TOEFL and TWE. It may be possible to use such a link to statistically adjust TWE scores according to topic difficulty.

In addition to topics, raters may also be a systematic

source of difficulty. It would be possible to calibrate the difference in relative severity of each rater at a particular reading by determining the disparity between the mean rating of rater A and the mean rating of rater B, with rater B being the mean second rating assigned by all other raters with whom rater A was paired. The difference from the norm in each rater's severity could be used to adjust the examinee's score. This approach provides a control over the raters only. It does not control for the topic, which is the greater source of concern among both test developers and raters. From an administrative perspective, one might question the wisdom of adjusting raters' scores instead of simply retraining the raters. Minor differences in rater severity will probably always exist as long as reliability is less than perfect, and it is questionable whether perfect rater reliability is even desirable, since it would suggest that superficial, easy-to-identify criteria were used as the basis for the rating.

Another approach to equating TWE scores is found in giving the examinee two different topics. Although costly as a regular operational procedure, such an approach might be tried as an experiment to determine if it is necessary to equate topics.

Even if it were established that it is desirable for topics to be equated, it is important to recognize that given the practical and financial implications of possible solutions, no

wholly satisfactory methodology exists to accomplish the task, at least in the traditional, statistical sense of the word equating,

Also, even if a suitable statistical methodology were available, it is still questionable whether TWE scores should be adjusted. From the beginning, the TWE program opted to employ a criterion-referenced scale. Within such a scale, each point has an agreed-upon meaning in terms of performance. Under such circumstances, it would seem inappropriate to alter a score through statistical equating. Perhaps two scores could be reported: a raw score based on the scoring guide and a scaled score representing topic difficulty. However, this might make score interpretation more difficult for score users. Should equating become possible, the issue will require careful consideration before a policy decision is made.

SUMMARY AND DISCUSSION

This long-term research agenda has outlined a very substantial program of research on the TWE. Although it was not the mandate of the project to establish specific priorities, some general recommendations have been made regarding the importance of various projects and areas of research at the time of this

writing. Of course, priorities may shift as certain projects are completed, and as new theoretical or operational concerns arise.

There is ample evidence that the TWE is a valid and reliable essay test. It was developed through careful research into the kinds of academic writing tasks required of foreign students at North American universities, and a major validity study was carried out before it becomes operational. Readers report positive reactions to the test, and universities are showing strong interest in recommending or requiring it of their nonnative English-speaking applicants for admission. Still, many of the concerns inherent in all essay tests require further research in the context of the TWE. In addition, some concerns unique to the TWE should be investigated.

There is also evidence of the construct validity of the TWE, because the test scores show no greater correlation with scores on other sections of TOEFL than those scores do with each other. Still, given the centrality of construct validity to a proper understanding of the meaning of a test score, the uniqueness (in comparison with other TOEFL scores) of the construct measured by this new component of the TOEFL merits further corroboration through research.

Currently, the matter of highest priority should be the question of the comparability of scores obtained on different

topics and different topic types. While this matter should be investigated as soon as possible, it may not be possible to satisfactorily address the issue without involving a large number of prompts. There is also a problem of controlling for differences in writing proficiency among different populations currently taking each prompt. Nevertheless such research would produce a better understanding of the comparability of scores.

The comparability question is also related to the issue of equating. Although no satisfactory method of equating exists at this time, ETS staff should continue to search for ways to equate TWE scores. Such methods might assist TWE program staff in improving the general quality of the test. However, even though equating is an important issue, it seems unlikely that a suitable and practical equating methodology will be encountered soon. Perhaps studies of equating methodologies should not be given high priority at this time.

At this point it appears that the enormous problems associated with the inter-rater reliability of essay tests have been largely overcome by the measures implemented for scoring the TWE. While inter-rater reliability indices are available for each administration thus far, none are available for the other types of reliability discussed in this paper: i.e., parallel-form reliability and the effect of different test conditions, such as time alloted for the response and permitting a choice of

topic. Research can also be carried out on raters. However, given the excellent inter-rater reliability obtained thus far, it is unlikely that much additional improvement can be obtained in the consistency of raters. Therefore, research on raters would seem to merit a lower priority than research on other types of reliability.

Because little scientific research has been conducted to date on the development of essay test prompts, research of this nature is desirable. The results should give test developers a better understanding of how to achieve desired ends. However, it may require considerable experience to make generalizations regarding specific phrases or approaches to the wording of essay prompts. Also, there is not clear agreement that the wording of a prompt is as important as other more basic characteristics, such as the topic itself. Therefore, while research on the wording of prompts could be useful, it need not be considered the highest priority at this time.

It seems that the most important issue facing the TWE program pertains to the empirical comparability of the skills tapped by the two topic types used on the test. While this is essentially a question of construct validity, the other broad areas of research outlined in this paper become more important as individual projects within them relate to this issue. Therefore, the comparability of topic types also relates to the need for

research on parallel-form reliability (since the topic types vary on different forms of the TWE) and on score equating (to determine if different topic types produce different scores). Although not a critical issue for the TWE program, test development research would be useful to test developers. At this time, lower priority should be given to studies of raters and studies of equating methodologies, since they do not relate directly to the topic type comparability issue.

If this research agenda is carried out, it is probable that much useful knowledge would be acquired. However, the agenda requires considerable time and money for successful completion. It also requires the efforts and careful concern of many involved in the TOEFL program, including TWE administrators, test developers, statistical analysts, and researchers, and the many non-ETS professionals who are involved with the TWE program in some way. Finally, the successful implementation of this agenda will require the support and attention of the TOEFL Research Committee and the TOEFL Policy Council. It is hoped the cooperation of all these groups can be obtained.

Notes

[1]This document was developed through a contract with the TOEFL Research Committee. A earlier version was submitted to the TOEFL Research Committee and approved as a final report in March 1988. The authors wish to thank the Committee for its support and advice regarding the manuscript.

[2]As described in the TOEFL Test and Score Manual (Educational Testing Service, 1987), the TOEFL program is governed by a Policy Council composed of 15 members representing the College Board, the Graduate Record Examinations Board, and other institutions and agencies, such as graduate schools of business, community colleges, nonprofit educational exchange agencies, and agencies of the United States government. The TOEFL Committee of Examiners and the TOEFL Research Committee are standing committees of the TOEFL Policy Council. The TWE Core Readers Group consists of seven writing specialists who prepare TWE topics and contribute to the management of the TWE program through advice and consultation. One member of the Core Readers Group serves on the TOEFL Committee of Examiners.

References

American Educational Research Association, American Psychological Association & National Council for Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second-language research. Language Learning, 32, 245-58.

Brown, J. D. & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. Language Learning, 34(4), 21-42.

Bridgeman, B. & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students (TOEFL Research Report 15). Princeton, NJ: Educational Testing Service.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of

English. TOEFL Research Report 19. Princeton, NJ: Educational Testing Service.

Conner, U. M. (1987, April). Understanding persuasive essay writing: Analytic versus holistic essay scoring--a place for both in ESL. Paper presented at the Fifth Annual Conference on Writing Assessment, Atlantic City, NJ.

Educational Testing Service. (1987). TOEFL Test and Score Manual. Princeton, NJ: Author.

Freedle, R. (1985, September). Language group differences in rhetorical writing patterns. Proposal submitted to the TOEFL Research Committee. Princeton, NJ: Educational Testing Service.

Greenberg, K. L. (1986). The development and validation of the TOEFL writing test: A discussion of TOEFL research reports 15 and 19. TESOL Quarterly, 20, 531-544.

Kaplan, R. B. (1972). The anatomy of rhetoric. Philadelphia, PA: Center for Curriculum Development.

Leonhardt, N. L. (1985). The effects of assigned versus open topics on the writing scores of university-level nonnative English speakers. Unpublished doctoral dissertation, Florida State University, Tallahassee.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. Educational Researcher, 10(9), 9-20.

Messick, S. (1987). Validity (Research Report RR-87-40). Princeton, NJ: Educational Testing Service.

Ross, J. & Stansfield, C. W. (March, 1987). Development of a long-term research agenda for the Test of Written English. Research precis submitted to the TOEFL Research Committee. Princeton, NJ: Educational Testing Service. (Internal document)

Roy, A.M. (1987, March). Evaluation of ESL students: The new TOEFL writing test and other standardized instruments. Paper presented at the postconvention workshop, Conference on College Composition and Communication, Atlanta, GA.

Stansfield, C. (1986a). A history of the Test of Written English: The developmental year. Language Testing, 3, 224-34.

Stansfield, C.W. (July 15, 1986b). Memorandum to R. Webster regarding observation of TWE. Princeton, NJ: Educational Testing Service. (Internal memorandum)

Stansfield, C. & Webster, R. (1986). The new TOEFL writing test. TESOL Newsletter, 20, 17-18.

Toulmin, S. E. (1958). The uses of argument. Cambridge: Cambridge University Press.

Wilson, K. M. (1982). A comparative analysis of TOEFL examinee characteristics (TOEFL Research Report 11). Princeton, NJ: Educational Testing Service.